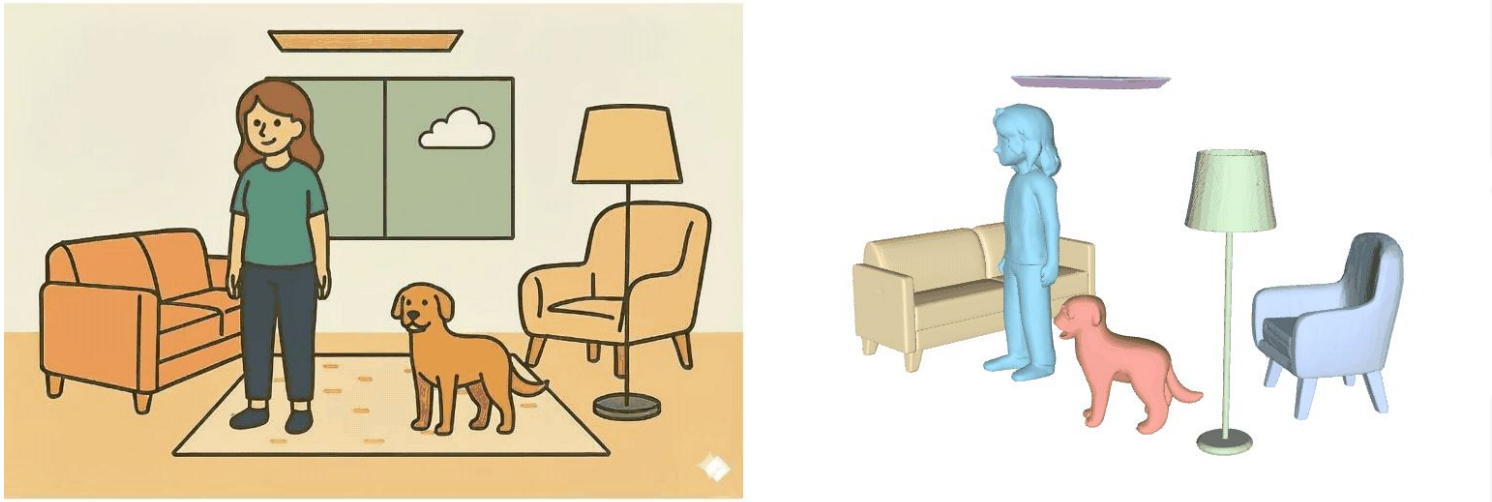
Inferring Compositional 4D Scenes without Ever Seeing One

Ahmet Berke Gokmen, Ajad Chhatkuli, Luc Van Gool, Danda Pani Paudel

CVPR 2026

INSAIT

Inferring Compositional 4D Scenes without Ever Seeing One



The Need

- Scene object decomposition is mainly tackled in static 3D settings.

PartCrafter [1]

They cannot move!

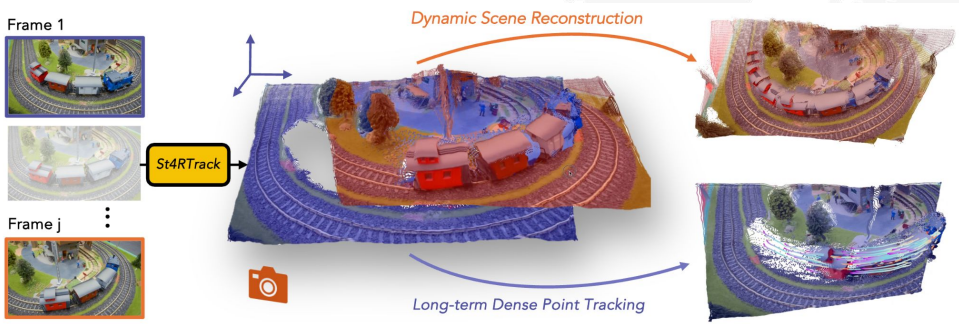
MIDI-3D [2]

SceneGen [3]

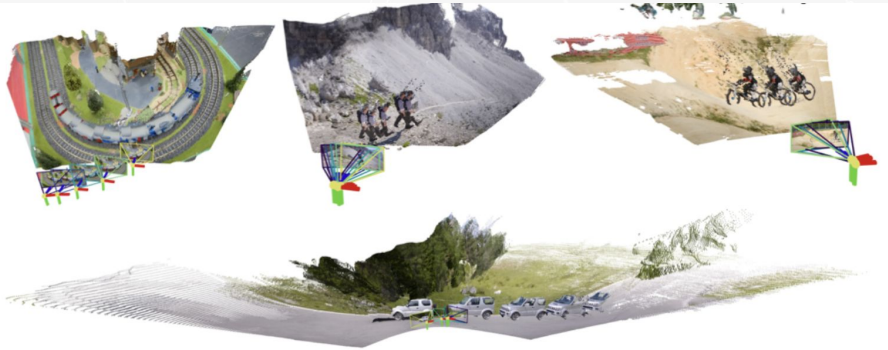
[1] Lin et al., PartCrafter, NeurIPS 2025
[2] Zehuan et al., MIDI-3D, CVPR 2025
[3] Meng et al., SceneGen, 3DV 2026

The Need

- Previous models do not explicitly reason about the occluded parts of each object given a single view observation because they were never trained to do so.



St4RTrack [1]



C4D [2]

Shape of Motion [3]

[1] Feng et al., St4RTrack, ICCV 2025
[2] Wang et al., C4D, ICCV 2025
[3] Wang et al., Shape of Motion, ICCV 2025

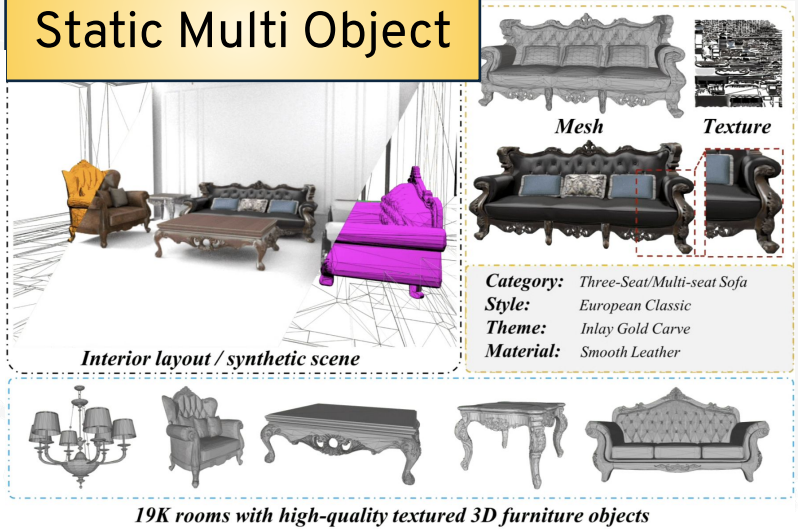
The Problem

- The are no “Compositional 4D” datasets as they are “expensive” to collect. We only have:

Dynamic Single Object

DeformingThings4D [1]

Static Multi Object

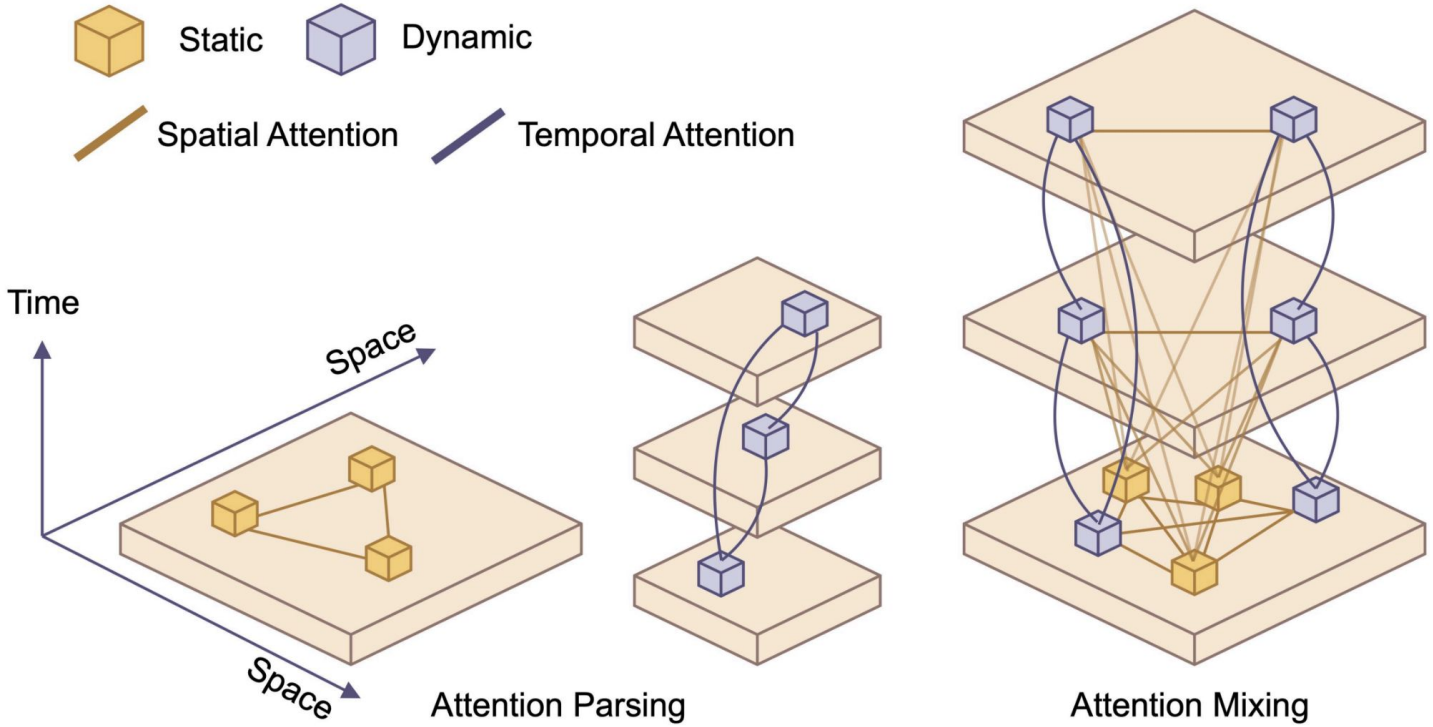


3D-Front [1]

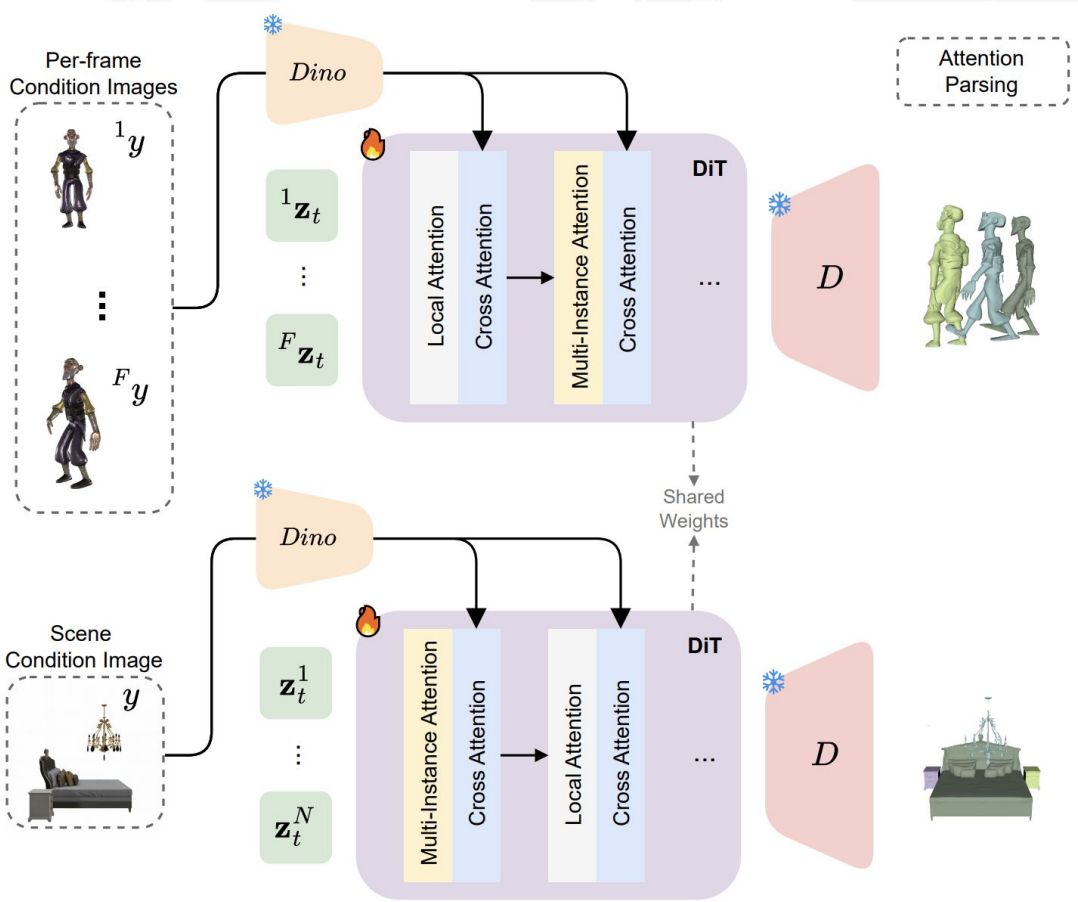
[1] Li et al., DeformingThings4D, ICCV 2021
[2] Fu et al., 3D-FRONT, arXiv 2021

The Idea

We can use priors from a pretrained single object generator and teach it temporal and spatial reasoning.



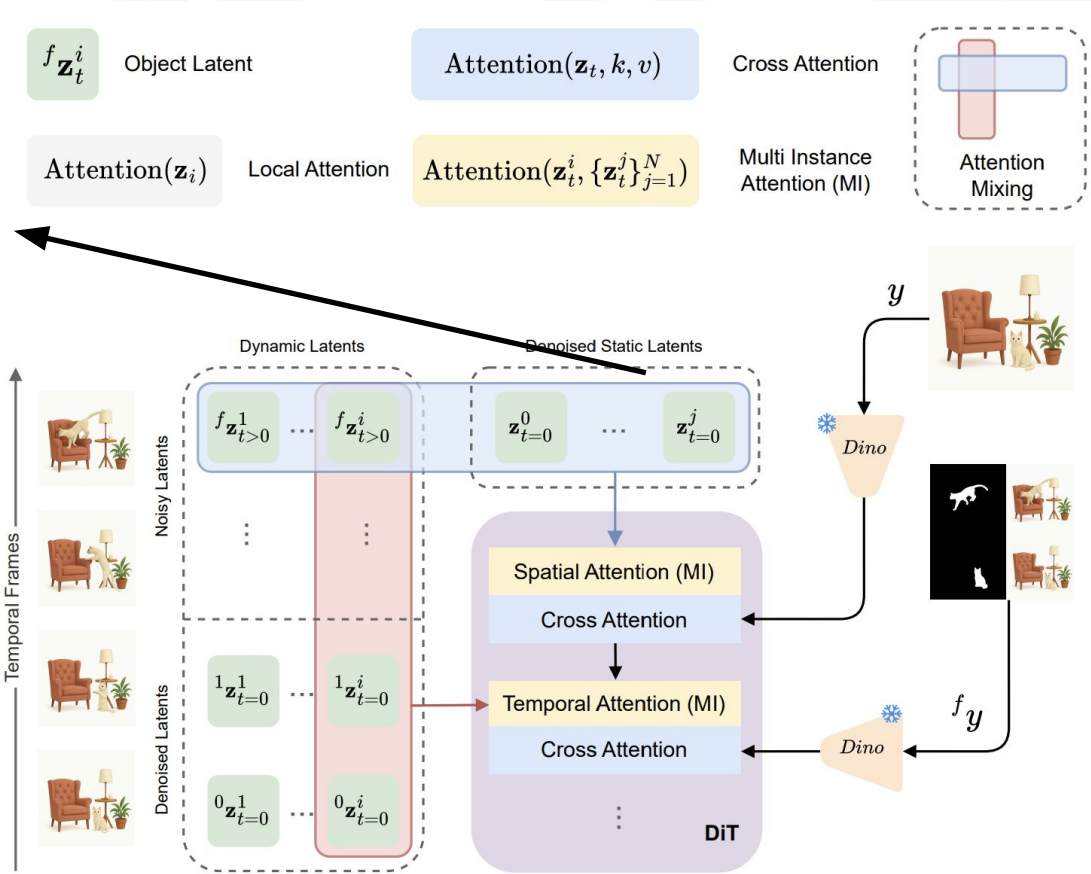
Training


Per-frame
Condition Images
Dino
Attention
Parsing
DiT
Local Attention
Cross Attention
Multi-Instance Attention
Cross Attention
D
Shared
Weights
Scene
Condition Image
Dino
DiT
Multi-Instance Attention
Cross Attention
Local Attention
Cross Attention
D

Inference (Mixing)

Wait, how does the model work when some of the latent tokens are all denoised and some are noisy?

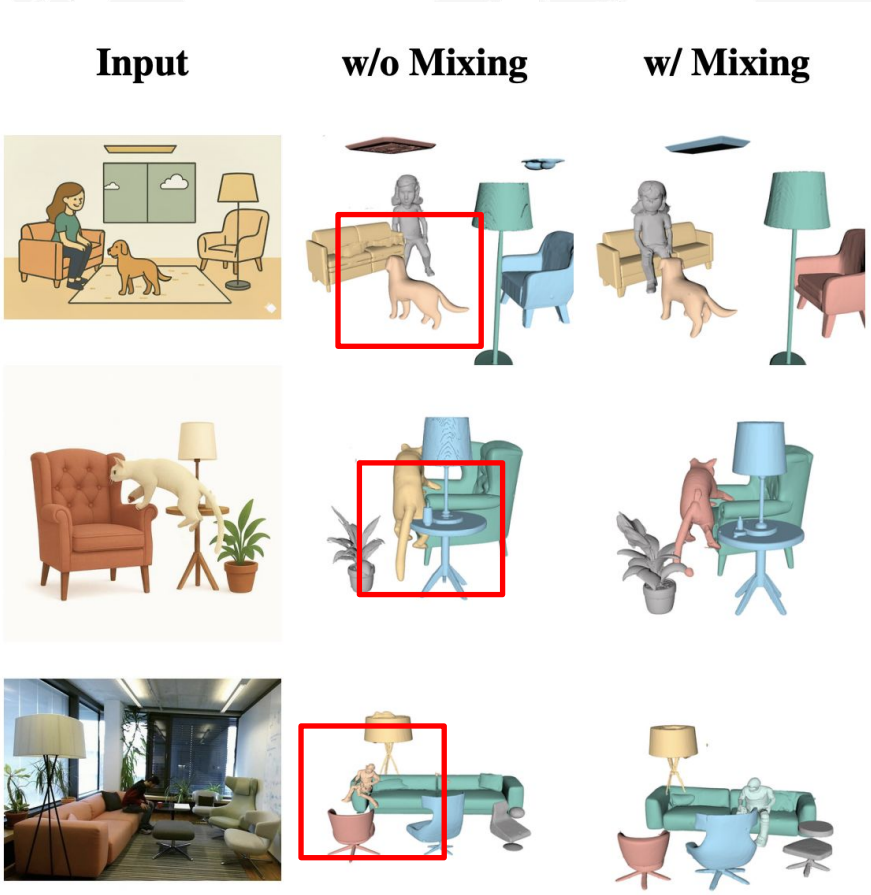
Diffusion Forcing!



The model was never trained on this type of input!

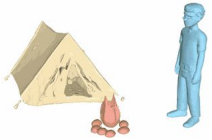
Why Do We Need Mixing?

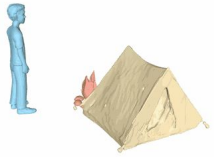
Why don't we just model static and dynamic objects separately and combine them in the same space?

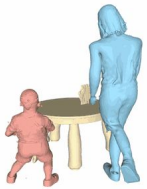
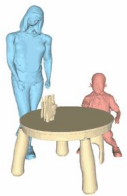
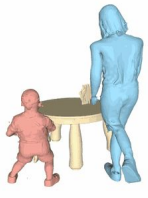
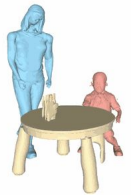


Without our "Mixing", the generation simply cannot figure out placements and relative scales of things.

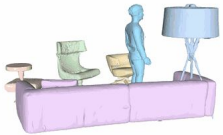
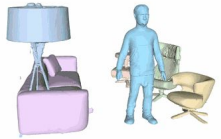
Additional Results

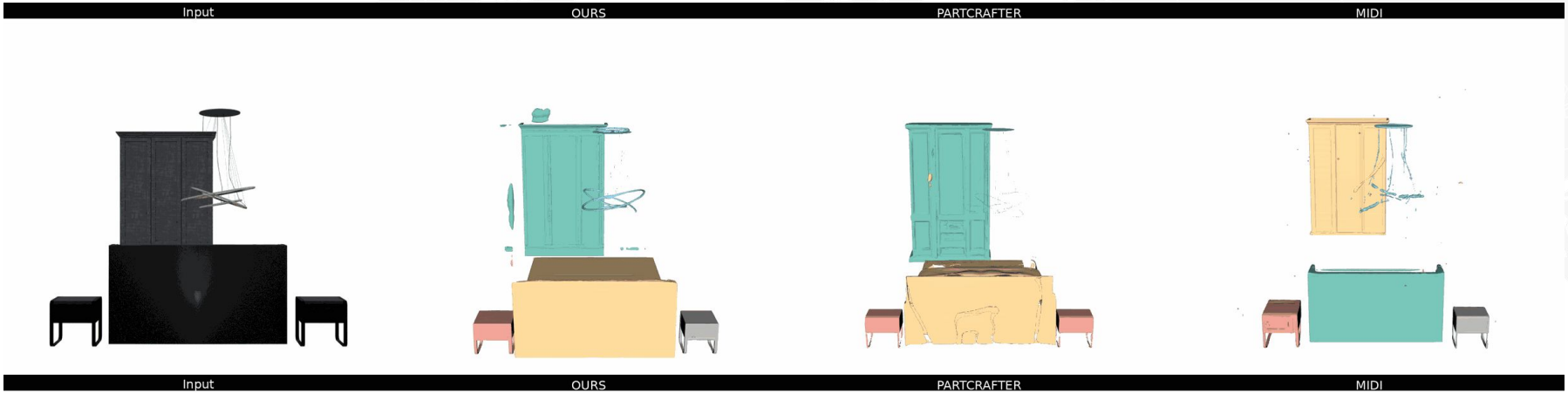


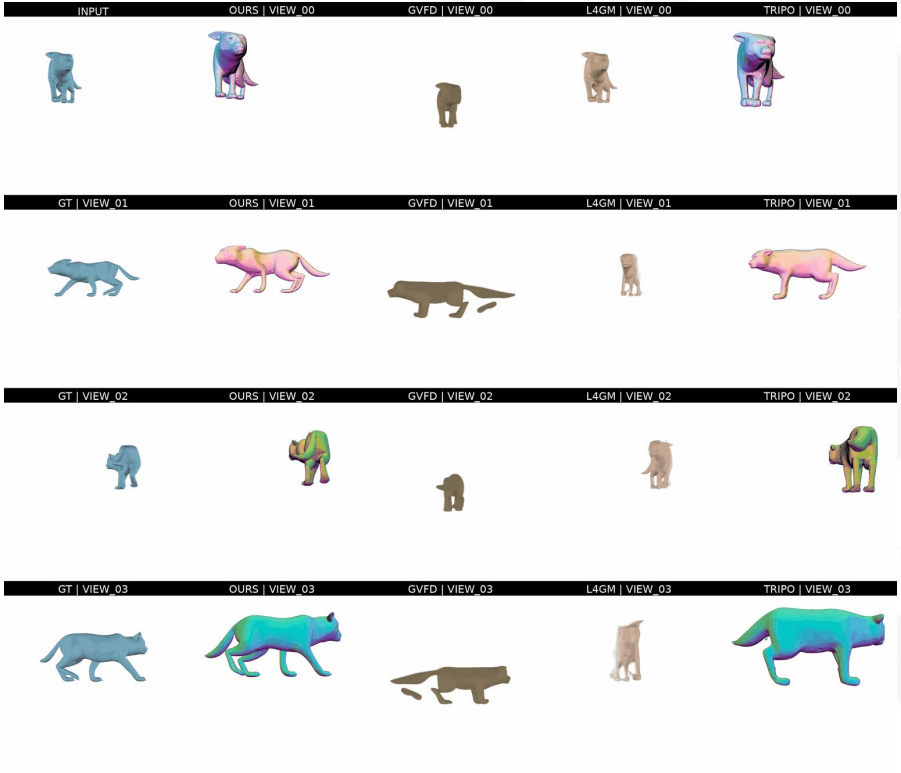


Additional Static 3D Compositional Results



Additional 4D Object Results



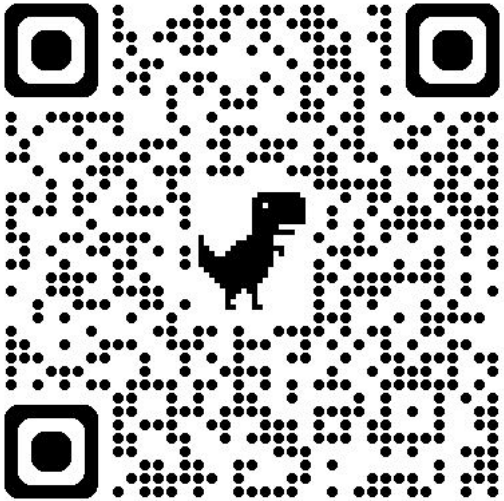
Quantitive Comparisons – 3D Scenes

Method	3D-FRONT [19]			3D-FRONT-Occluded [19]		Runtime
	CD ↓	F-Score ↑	IoU ↓	CD ↓	F-Score ↑	Seconds ↓
MIDI [29]	*0.1445*	*0.7829*	0.0027	*0.1781*	0.7387	3.51
PartCrafter [47]	0.1751	0.7569	**0.0017**	0.1951	*0.7461*	**2.60**
Ours	**0.0909**	**0.8069**	*0.0018*	**0.1256**	**0.7521**	*2.63*

Quantitive Comparisons – 4D Objects

Method	DeformingThings [42]			Objaverse [15]			Runtime
	CD ↓	F-Score ↑	IoU ↑	CD ↓	F-Score ↑	IoU ↑	Sec. ↓
V2M4 [10]	0.1678	0.4759	0.1533	0.1595	0.4903	0.2135	61.98
L4GM [68]	0.2633	0.3167	—	0.2308	0.3236	—	**1.22**
GVFD [94]	0.2806	0.2706	—	0.2728	0.2856	—	2.87
TripoSG [43]	*0.1558*	*0.5179*	*0.1784*	**0.1107**	*0.6585*	*0.2874*	*2.59*
Ours	**0.1144**	**0.8388**	**0.4191**	*0.1205*	**0.7349**	**0.3413**	4.48

Thanks!



Visit our website:
https://com4d.insait.ai/